MPI on BG/L at ANL

Rusty Lusk
Mathematics and Computer Science Division
Argonne National Laboratory

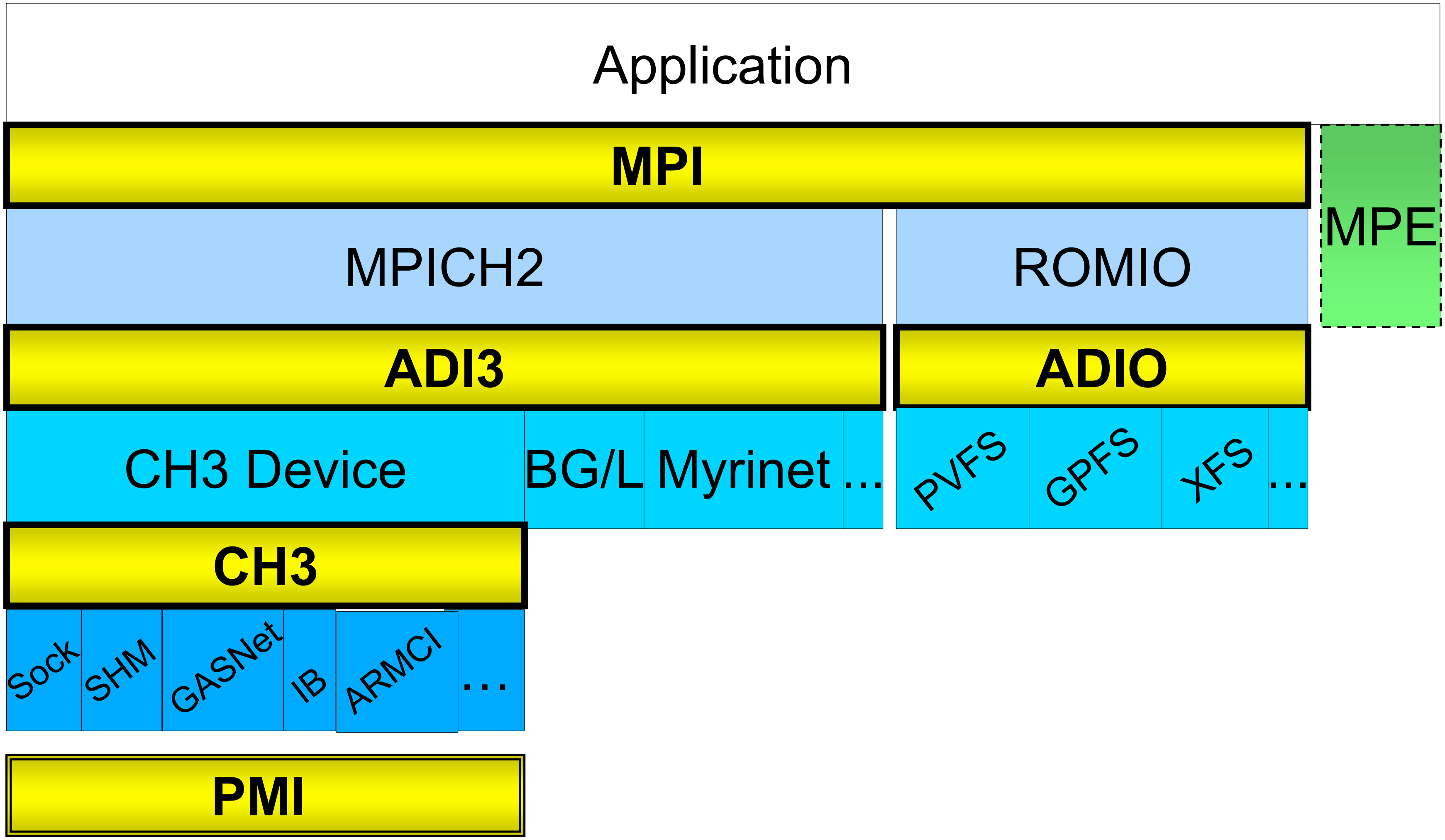
Outline

- MPI, MPI-2, and MPICH2
- Some basic experimental results on MPI performance
- A short look at a performance tool on BG/L
- Some near-term relevant ANL projects
- Some longer-term potential collaborative projects

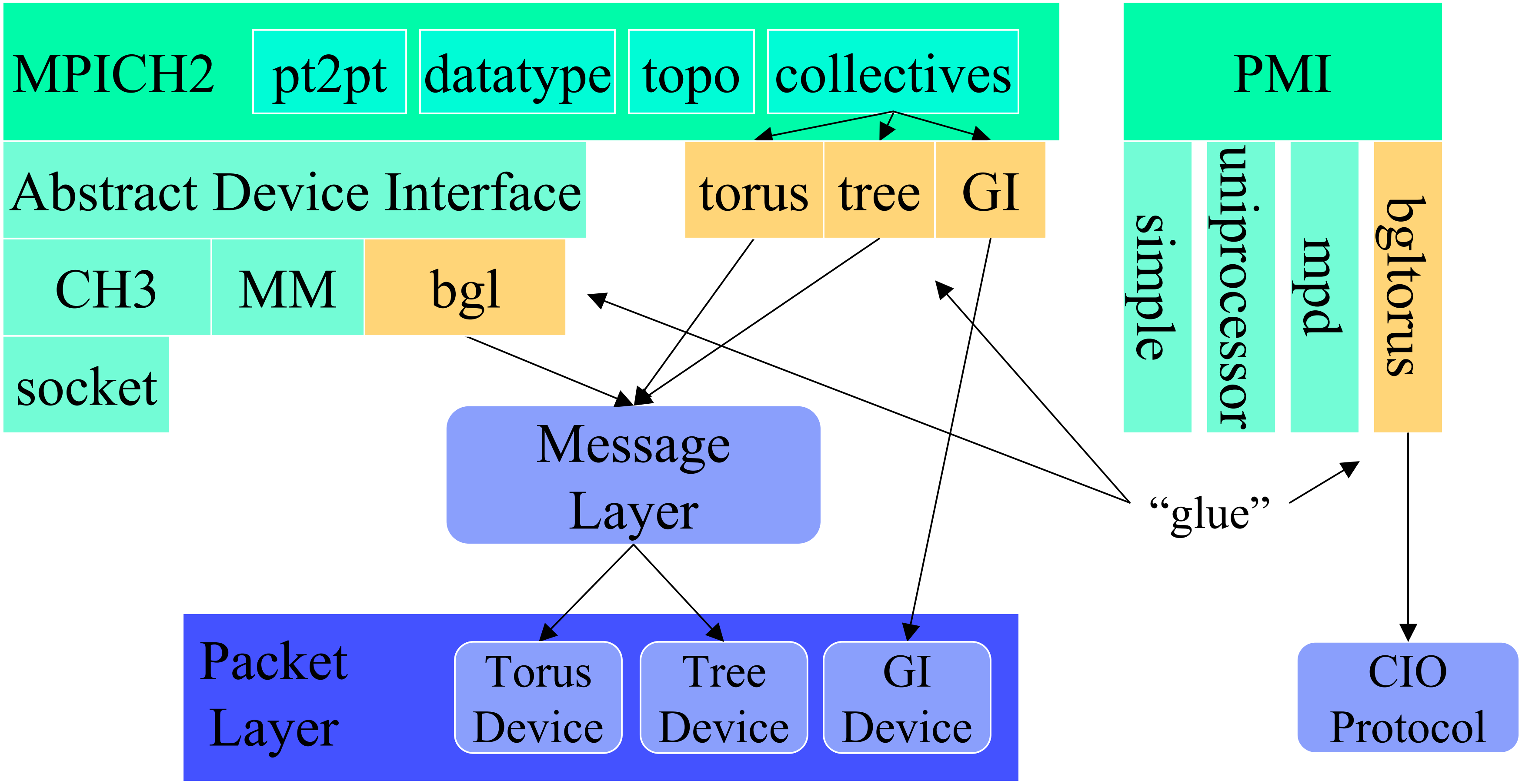
MPI Implementation at ANL and IBM

- MPICH2 is an all-new, open-source, portable implementation of the full MPI-2 standard
- Several vendors are using it as the basis of their MPI implementations.
- IBM and ANL have collaborated on using MPICH2 as the basis of BG/L’s MPI (MPI-1, so far)

MPICH2 Structure



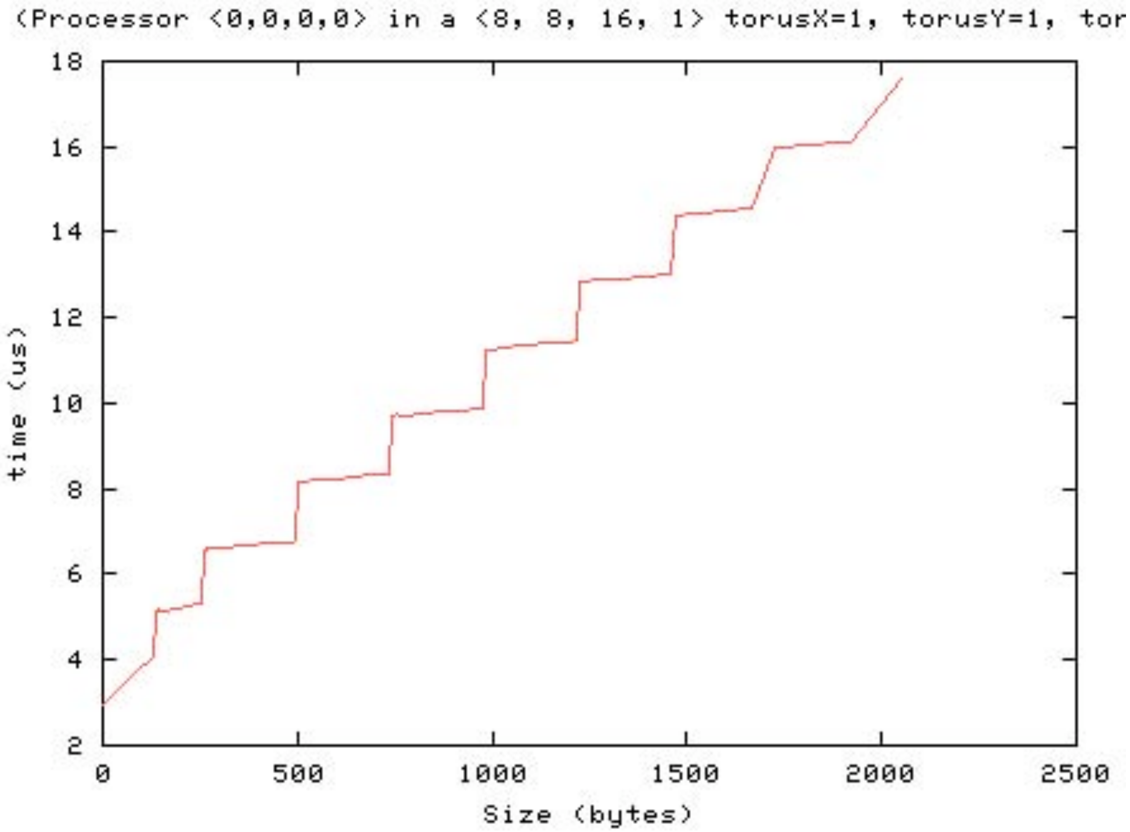
IBM BG/L MPI Software Architecture



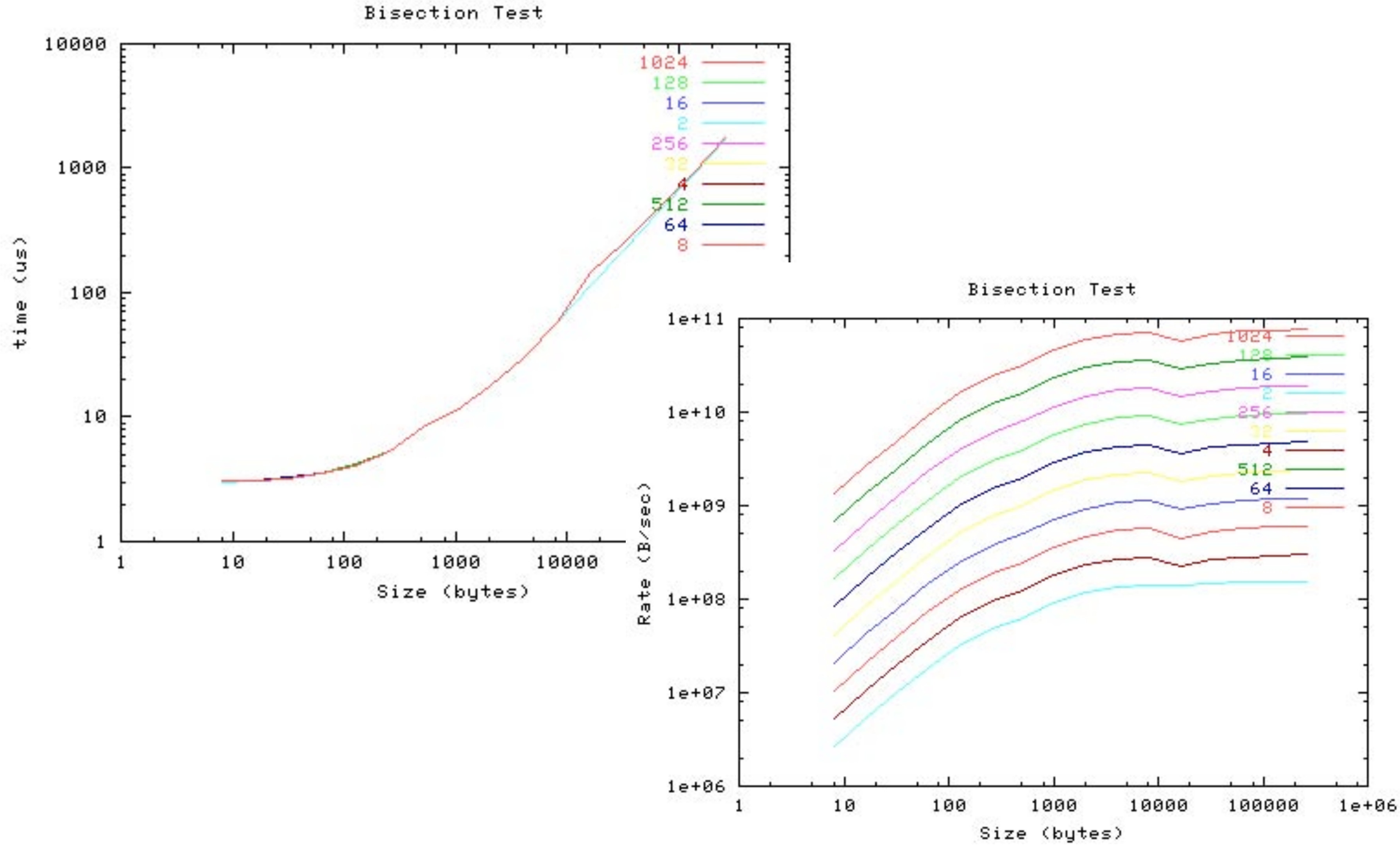
This slide courtesy of IBM

Performance Summary

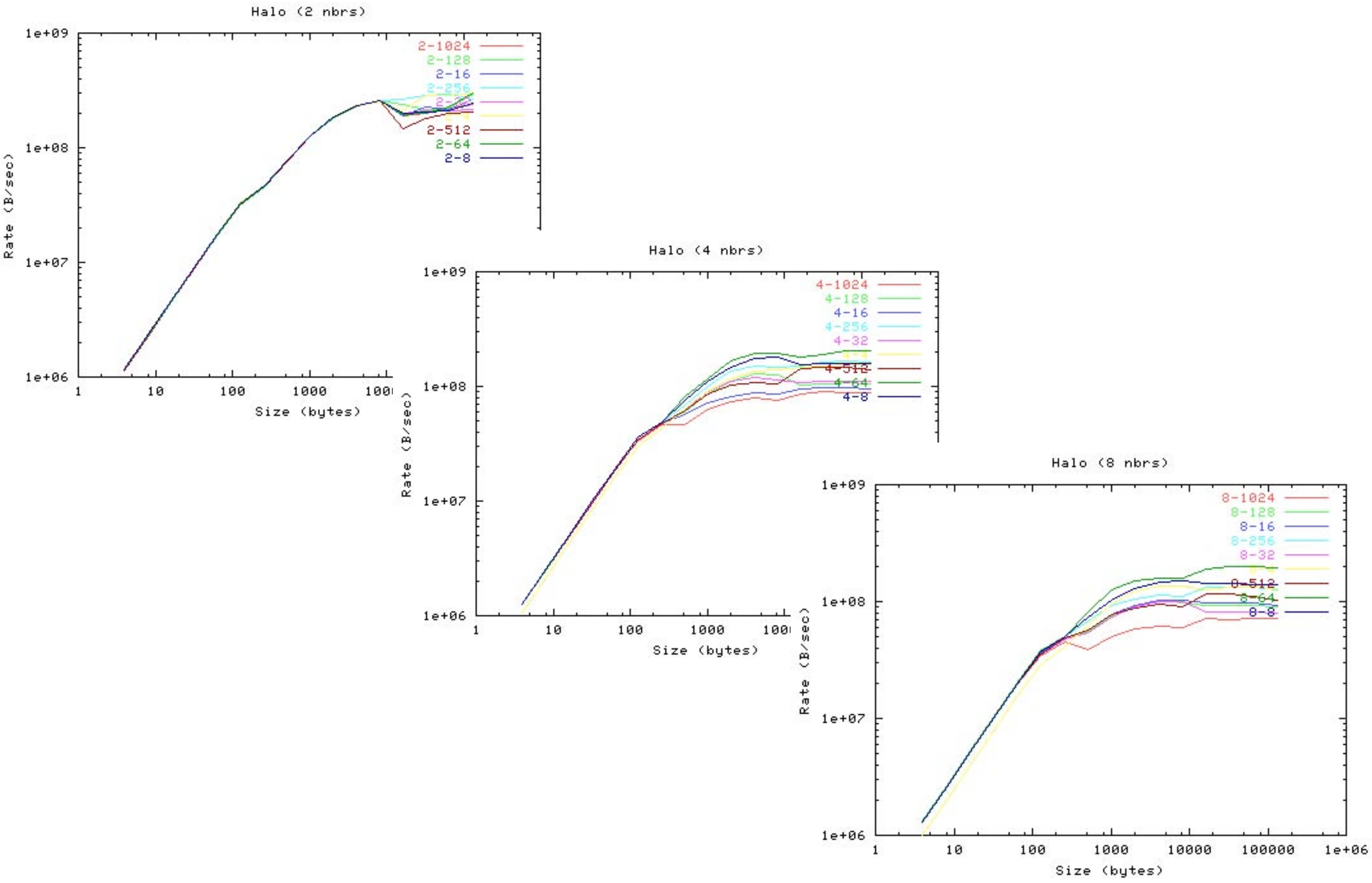
- Recent tests on Argonne’s one-rack machine
- From http://www.mcs.anlo.gov/~gropp/projects/parallel/BGL/mpptest (other tests nearby)
- Latency:



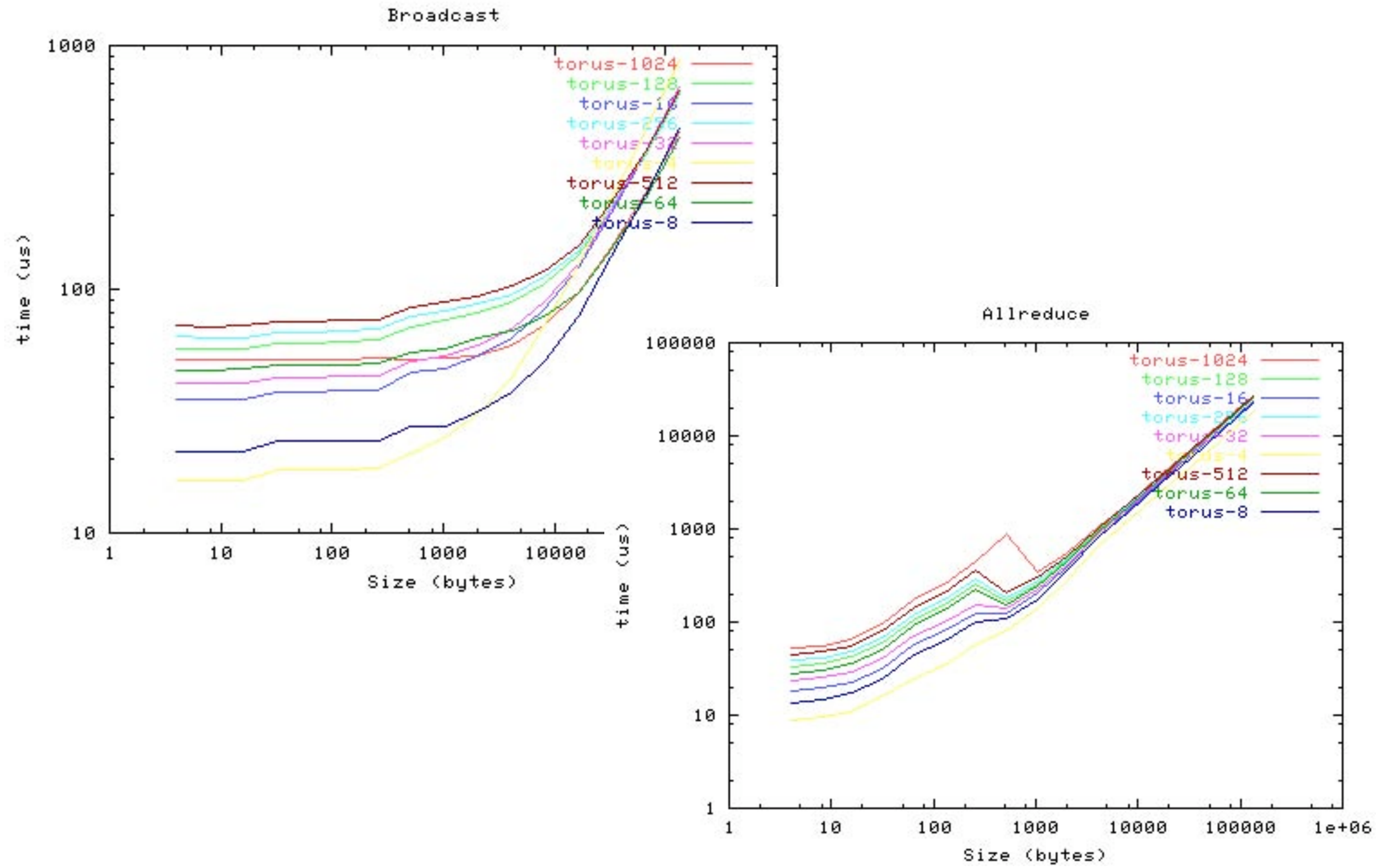
Bisection Performance



Halo Exchange



Collective Operations

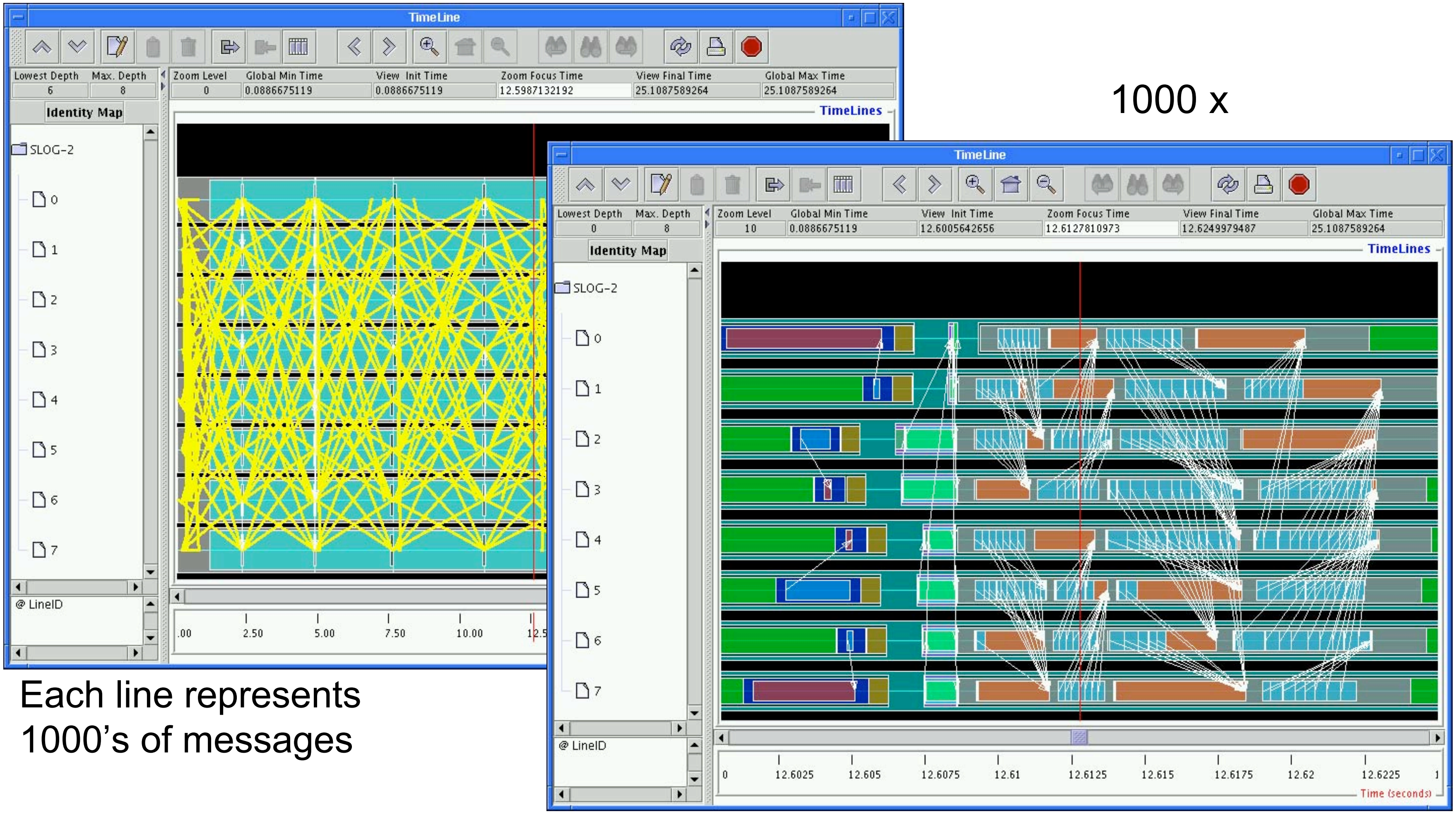


MPI Performance Summary

- With our own tests on our own machine, BG/L’s MPI is reliable, fast, consistent, and scalable.

Jumpshot on BG/L

1000 x



Lowest Depth	Max. Depth	Zoom Level	Global Min Time	View Init Time	Zoom Focus Time	View Final Time	Global Max Time
6	8	0	0.0886675119	0.0886675119	12.5987132192	25.1087589264	25.1087589264

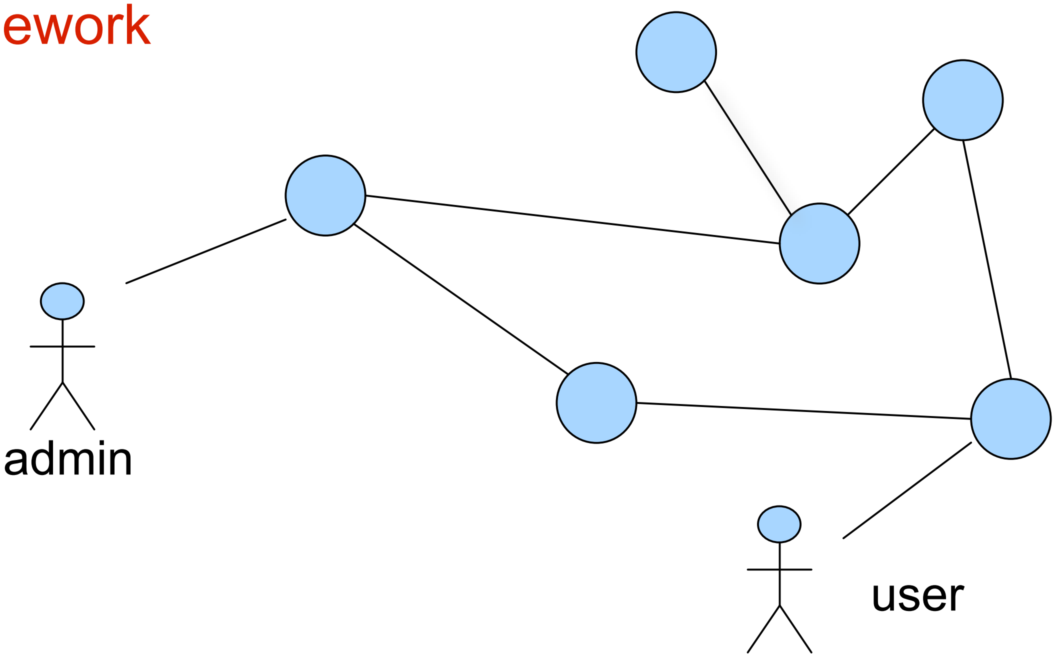
Lowest Depth	Max. Depth	Zoom Level	Global Min Time	View Init Time	Zoom Focus Time	View Final Time	Global Max Time
0	8	10	0.0886675119	12.6005642656	12.6127810973	12.6249979487	25.1087589264

Each line represents 1000's of messages

Detailed view shows opportunities for optimization

The System Software Environment

- We are participants in the Scalable Systems Software SciDAC project, which has developed a component architecture for system software
- Whole suite currently running on Chiba City cluster
- Some components being ported to BG/L environment
  - Communication framework
  - Process manager
  - Queue manager
  - Scheduler



Near-Term MPI Projects

- MPICH2 release 1.0.1 (this week) has slots for specialized topology routines, analogous to slots for specialized collective routines. We plan to work with IBM to ensure that this approach enables new optimizations.
  - Should help applications tune themselves
- We plan to conduct research and develop new MPICH collective routines that base their algorithms on the topology routines.
- Incorporating new optimized datatype handling
- MPICH2 supports the MPI standard **`mpiexec`** with several useful extensions (such as passing environment variables). We hope to merge our approach with IBM's **`mpirun`**, which is a work in progress.
- Porting some system software components
  - Integrating with existing IBM system software

Longer-term Collaborative MPI Projects

- MPI-2 (already in MPICH2)
  - MPI-I/O to a fast parallel file system
    - Rob’s talk: need better language for implementing ROMIO on compute nodes
  - One-sided operations
    - We currently are working with a neuroscience application that is a good match to MPI_Put/Get
    - Short-term: port MPICH2 version
    - Long-term: customized for BG/L
- System software
  - Several ways to improve the environment seen by users being explored

The End

PLAN GLIMB

Rusty Lusk
Mathematics and Computer Science Division
Argonne National Laboratory